L-DiffER: Single Image Reflection Removal with Language-based Diffusion Model (Supplementary Material)

Yuchen Hong[1,2#] Haofeng Zhong[1,2,3#] Shuchen Weng[1,2]
Jinxiu Liang[1,2] Boxin Shi[1,2,3*]

[1] State Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University
[2] National Engineering Research Center of Visual Technology, School of Computer Science, Peking University
[3] AI Innovation Center, School of Computer Science, Peking University
{hfzhong, shuchenweng, cssherryliang, shiboxin}@pku.edu.cn, yuchenhong.cn@gmail.com

In the supplementary material, we introduce details about the time-variant coefficients in the proposed iterative condition refinement strategy, show evaluations on reflection recovery, and provide ablation studies on the network design, the saturation-aware structure constraint, loss functions, and language descriptions. We further conduct comparisons on the model size and inference time, alongside additional qualitative comparisons with state-of-the-art reflection removal methods.

6 Details about the time-variant coefficients

In this section, we provide details of time-variant coefficients β_t and γ_t ($t \in \{1, ..., T\}$) in the proposed iterative condition refinement strategy, controlling the refinement of the structure and color condition in Eq. (4) (corresponding to footnote 2 in the main paper). We set β_t and γ_t to the same value for synchronously updating the structure and color condition, which are defined as follows:

$$\beta_t = \gamma_t \coloneqq \prod_{i=1}^{t}(1 - \psi_i), \tag{17}$$

where the variants $\psi_i \in \Psi \coloneqq \{\psi_1, ..., \psi_T\}$ are constants increasing linearly from $\psi_1 = 10^{-4}$ to $\psi_T = 0.02$ inspired by [3], which is reasonable since these constants are small enough related to data scaled to $[-1, 1]$, guaranteeing that the forward and reverse processes exhibit nearly the same functional form.

7 Evaluations on reflection recovery

In this section, we conduct experiments to evaluate the recovery of reflection layers (corresponding to footnote 4 in the main paper). Quantitative comparisons

# Equal contributions. * Corresponding author.

Table 3: Comparison of quantitative results on real datasets for evaluating the recovery of reflection layers, compared with several state-of-the-art single-image reflection removal methods [2,9,10,12,17,21,23]. ↑ (↓) indicates larger (smaller) values are better. **Bold** numbers indicate the best-performing results.

Dataset (size)	Metric	Method							
		Zhang *et al.*	CoRRN	ERRNet	IBCLN	Dong *et al.*	YTMT	DSRNet	Ours
SIR^2 (500)	PSNR↑	24.64	25.94	25.37	25.76	26.07	24.48	26.24	**26.45**
	SSIM↑	0.486	0.513	0.506	0.502	0.519	0.479	0.521	**0.532**
Real20 (20)	PSNR↑	25.16	24.83	25.41	25.29	25.21	25.09	25.36	**25.51**
	SSIM↑	0.506	0.487	0.519	0.508	0.532	0.515	0.529	**0.548**
Nature (20)	PSNR↑	24.97	24.73	24.25	25.55	25.65	24.51	24.89	**25.88**
	SSIM↑	0.511	0.497	0.542	0.552	0.568	0.458	0.561	**0.572**
Average (540)	PSNR↑	24.67	25.85	25.33	25.73	26.02	24.50	26.16	**26.39**
	SSIM↑	0.488	0.511	0.508	0.504	0.521	0.480	0.523	**0.534**

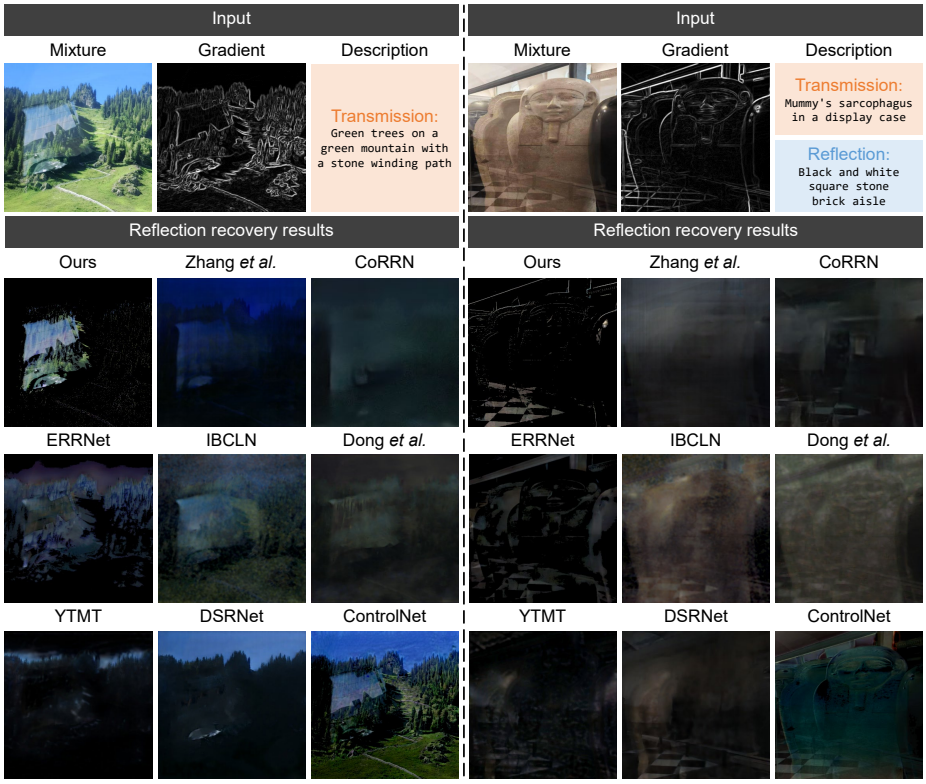

Fig. 8: Qualitative comparison of estimated reflection layers on real mixture images collected from the Internet, compared with several single-image methods [2, 9, 10, 12, 17, 21, 23] and a diffusion-based method ControlNet [22]. Please zoom in for details.

Table 4: Ablation studies on the network design and the loss function. ↑ (↓) indicates larger (smaller) values are better. **Bold** numbers indicate the best-performing transmission recovery results.

Metric	Abl. on network design		
	Both cond. enc.	Single trainable	Ours
PSNR↑	24.39	24.86	**25.08**
SSIM↑	0.889	0.896	**0.905**
LPIPS↓	0.141	0.135	**0.123**
NIQE↓	4.696	4.678	**4.322**
FID↓	57.12	52.48	**46.58**

are conducted on existing reflection removal datasets (*i.e.*, SIR^2 [18], Real20 [23], and Nature [12]), and ground truths of reflection layers are obtained as in [10]. We employ PSNR [11] and SSIM [20] as error metrics following [4–8, 24]. Qualitative comparisons are conducted on images from the Internet. We compare the proposed method with state-of-the-art single-image methods (including Zhang *et al.* [23], CoRRN [17], ERRNet [21], IBCLN [12], Dong *et al.* [2], YTMT [9], and DSRNet [10]) and a language-based diffusion model ControlNet [22]. As ERRNet [21], ControlNet [22], and the proposed method only estimate the transmission layers $\widetilde{\mathbf{T}}$ from the mixture images $\mathbf{M}$ by networks, we obtain reflection layers $\widetilde{\mathbf{R}}$ by $\widetilde{\mathbf{R}} = \mathbf{M} - \widetilde{\mathbf{T}}$ following [15, 17]. As quantitative and qualitative results shown in Table 3 and Fig. 8 (corresponding to Fig. 5 in the main paper), contributing to the high-fidelity and faithful recovery of transmission layers, the proposed method achieves the state-of-the-art performance on reflection recovery, which indicates the efficacy of the proposed iterative condition strategy and multi-condition constraint mechanism for leveraging the language-based diffusion model.

8 Additional ablation studies

In this section, we conduct ablation studies to investigate the effectiveness of the network design, the saturation-aware structure constraint, loss functions, and language descriptions (corresponding to footnote 5 in the main paper).

Ablation studies on the network design. As mentioned in Sec. 3.1 of the main paper, the proposed L-diffER utilizes a color encoder $\mathcal{E}^{\mathrm{c}}$ and a structure encoder $\mathcal{E}^{\mathrm{s}}$ to transform the color and structure condition into latent space, respectively. In practice, the color encoder $\mathcal{E}^{\mathrm{c}}$ uses the compression encoder of Stable Diffusion (SD) [13] and the structure encoder $\mathcal{E}^{\mathrm{s}}$ uses the condition encoder of ControlNet [1, 22]. We conduct an ablation study by replacing $\mathcal{E}^{\mathrm{c}}$ with the condition encoder of ControlNet [22] (denoted as 'Both cond. enc.') that is also the network architecture of $\mathcal{E}^{\mathrm{s}}$ to investigate the influence of encoders. Furthermore, following ControlNet [22], two trainable copied modules (denoted by the pink blocks in Fig. 2 of the main paper) of SD [13] are employed to separately extract color and structure features from color and structure latents, so we conduct another ablation study by only using a single trainable copied module to jointly extract color and structure features (denoted as 'Single trainable'). As

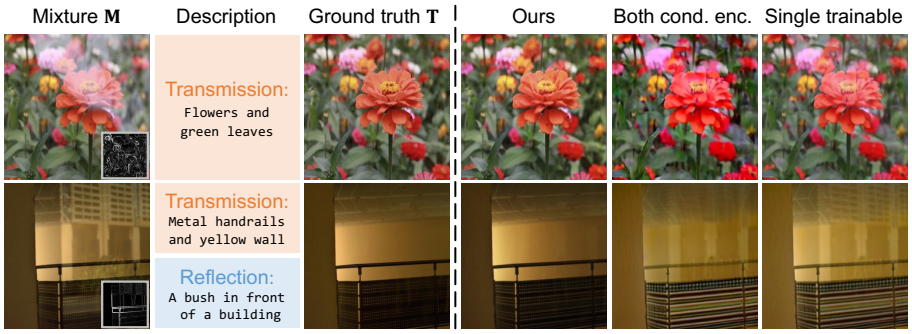

Fig. 9: Ablation study on the network design. We show the gradient map (*i.e.*, the original structure condition) at the lower right of each mixture image (*i.e.*, the original color condition). Please zoom in for details.

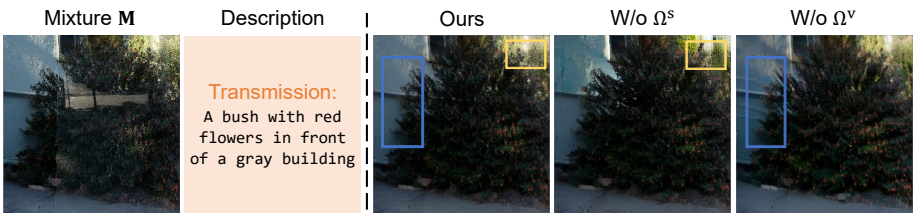

Fig. 10: The effect of $\mathbf{\Omega}^{\mathrm{s}}$ and $\mathbf{\Omega}^{\mathrm{v}}$. Please zoom in for details.

Table 5: Ablation studies on pixel loss functions. ↑ (↓) indicates larger (smaller) values are better. **Bold** numbers indicate the best-performing transmission recovery results.

Method	PSNR↑	SSIM↑	LPIPS↓	NIQE↓	FID↓
W/o $\mathcal{L}_{\mathrm{RGB}}$	24.87	0.896	0.137	4.702	52.32
W/o $\mathcal{L}_{\mathrm{grad}}$	24.68	0.894	0.140	4.813	53.47
W/o $\mathcal{L}_{\mathrm{num}}$	24.97	0.900	0.131	4.593	48.64
W/o $\mathcal{L}_{\mathrm{pix}}$	24.22	0.884	0.149	4.751	58.21
Ours	**25.08**	**0.905**	**0.123**	**4.322**	**46.58**

results shown in Table 4 and Fig. 9, modifications on the encoder or the trainable copied module lead to performance degradation, indicating the effectiveness of our network design for condition injection.

Ablation studies on the saturation-aware structure constraint. As mentioned in the main paper, $\mathbf{\Omega}^{\mathrm{s}}$ in Eq. (7) is designed to indicate saturated regions for retaining the generative capability of the diffusion model in these regions, and $\mathbf{\Omega}^{\mathrm{v}}$ in Eq. (8) is to indicate valid edges for preventing overly generation in non-saturated regions. We conduct ablation studies shown in Fig. 10 by removing $\mathbf{\Omega}^{\mathrm{s}}$ and $\mathbf{\Omega}^{\mathrm{v}}$, respectively, which verifies the effectiveness of the two components.

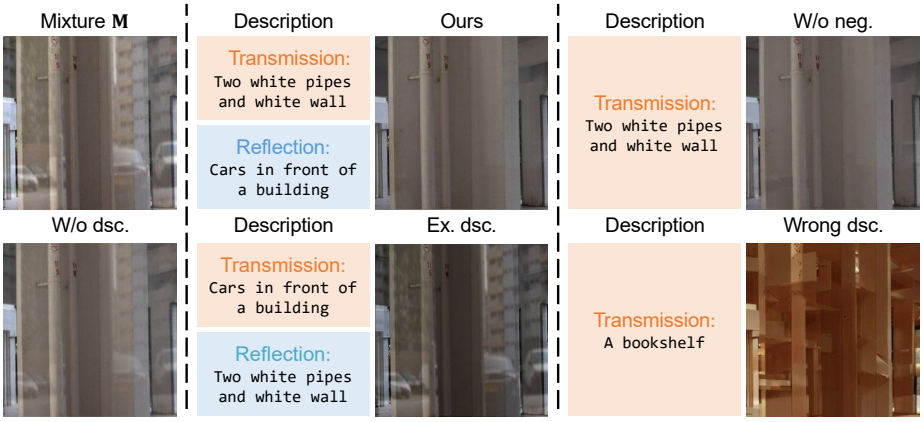

Fig. 11: The effect of input language descriptions. Please zoom in for details.

Ablation studies on loss functions. We further investigate the effect of the loss functions (introduced in Sec. 3.4 of the main paper) as shown in Table 5. Compared with the complete model, disabling the RGB loss ('W/o $\mathcal{L}_{\text{RGB}}$') and gradient loss ('W/o $\mathcal{L}_{\text{grad}}$') degrade performance significantly since inaccurate conditions mislead transmission recovery. Disabling the numerical loss ('W/o $\mathcal{L}_{\text{num}}$') also causes performance decline, suggesting the efficacy of the supervision on images' mean and variance. Besides, the performance reduction of the variant 'W/o $\mathcal{L}_{\text{pix}}$' demonstrates the necessity of conducting supervision on the pseudo transmission layers $\mathbf{T}_{0|t}$ at the pixel level. Note that we do not disable $\mathcal{L}_{\text{ldm}}$ since diffusion models can not train without it.

Ablation studies on language descriptions. We conduct ablation studies on the numbers and the order of input language descriptions to verify the effectiveness of language guidance. As shown in Fig. 11, abandoning the description of the reflection layer (W/o neg.) causes a few reflection residuals in the example, and more reflections remain if directly using an empty description (W/o dsc.). If exchanging the order of prompts (Ex. dsc.), results will degrade due to different statistics of transmission and reflection layers [14,16], but the contents of recovered results still conform to the prompt. Similarly, using the wrong description of the transmission layer (Wrong dsc.) will cause a completely different recovered result. By utilizing both language descriptions of two layers, the proposed method achieves high-fidelity reflection removal, which indicates the efficacy of introducing language descriptions.

9 Comparisons of the model size and inference time

In this section, we show the comparisons of the model size and inference time in Table 6. The image size is 384×384, and we run the inference on an NVIDIA GeForce RTX 3090. Though owning more parameters and more inference time

Table 6: Comparisons of the model size and inference time. We show both trainable and total parameters for diffusion-based methods.

Metric	Single-image				Diffusion-based	
	CoRRN [57]	IBCLN [30]	YTMT [20]	DSRNet [21]	ControlNet [66]	Ours
Params	59.5M	21.6M	73.4M	137.6M	364.4M/1.4B	728.8M/1.8B
Time (s)	0.079	0.127	0.206	0.361	11.547	12.685

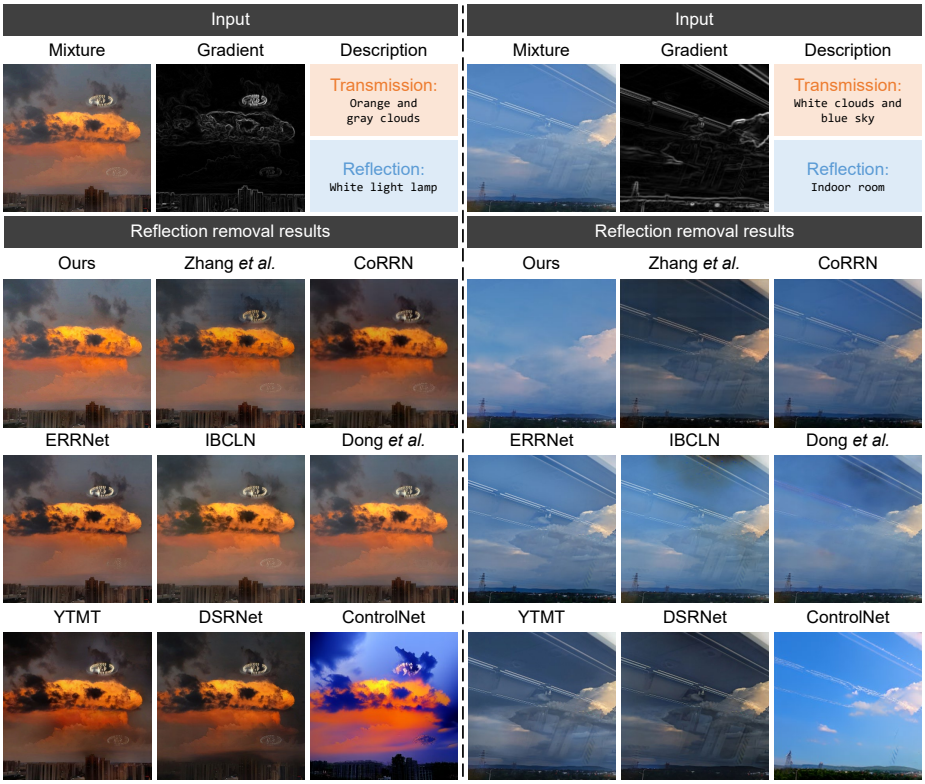

Fig. 12: Qualitative comparison of estimated transmission layers on real mixture images collected from the Internet, compared with several single-image methods [2, 9, 10, 12, 17, 23] and a diffusion-based method ControlNet [22]. Please zoom in for details.

than traditional single-image methods, it is worth noting that the proposed method pioneers in introducing large language-based diffusion models for reflection removal to solve the most concerned problems, *i.e.*, relieving the ill-posedness and tackling low-transmitted or saturated reflections, which facilitates future research in a new perspective.

Fig. 13: High-resolution real examples with low-transmitted or saturated reflections. Please zoom in for details.

10 Additional qualitative results

In this section, additional qualitative experiments are conducted on real images to show the effectiveness of the proposed language-based diffusion model for reflection removal. Experimental settings are the same as in Sec. 4.1 of the main paper, and qualitative results are shown in Fig. 12. As can be observed, single-image reflection removal methods [2, 9, 10, 12, 17, 21, 23] fail in low-transmitted reflection regions, and ControlNet [22] generates results with color shifts and structure distortions. The proposed method thoroughly removes reflections and recovers clear transmission layers even in low-transmitted and saturated reflection regions (*e.g.*, the reflections of white lights in the left and right example of Fig. 12), which demonstrates the effectiveness of using language descriptions to provide auxiliary semantic information and the unique advantage of leveraging generative priors from diffusion models [13] by employing the proposed refinement strategy and constraint mechanism for conditions.

We further present two real examples with low-transmitted or saturated reflections. By adopting a patch aggregation method [19], we can obtain their corresponding high-resolution (*i.e.*, 1024×1024) results shown in Fig. 13, which verifies the robustness of the proposed method.

References

1. Chang, Z., Weng, S., Zhang, P., Li, Y., Li, S., Shi, B.: L-cad: Language-based colorization with any-level descriptions using diffusion priors. In: Proc. of Advances in Neural Information Processing Systems (2023)
2. Dong, Z., Xu, K., Yang, Y., Bao, H., Xu, W., Lau, R.W.: Location-aware single image reflection removal. In: Proc. of International Conference on Computer Vision (2021)
3. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Proc. of Advances in Neural Information Processing Systems (2020)
4. Hong, Y., Chang, Y., Liang, J., Ma, L., Huang, T., Shi, B.: Light flickering guided reflection removal. International Journal of Computer Vision (2024)
5. Hong, Y., Lyu, Y., Li, S., Cao, G., Shi, B.: Reflection removal with nir and rgb image feature fusion. IEEE Transactions on Multimedia (2022)
6. Hong, Y., Lyu, Y., Li, S., Shi, B.: Near-infrared image guided reflection removal. In: Proc. of International Conference on Multimedia and Expo (2020)

7. Hong, Y., Zheng, Q., Zhao, L., Jiang, X., Kot, A.C., Shi, B.: Panoramic image reflection removal. In: Proc. of Computer Vision and Pattern Recognition (2021)
8. Hong, Y., Zheng, Q., Zhao, L., Jiang, X., Kot, A.C., Shi, B.: PAR^{2}Net: End-to-end panoramic image reflection removal. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
9. Hu, Q., Guo, X.: Trash or treasure? an interactive dual-stream strategy for single image reflection separation. Proc. of Advances in Neural Information Processing Systems (2021)
10. Hu, Q., Guo, X.: Single image reflection separation via component synergy. In: Proc. of International Conference on Computer Vision (2023)
11. Huynh-Thu, Q., Ghanbari, M.: Scope of validity of psnr in image/video quality assessment. Electronics letters (2008)
12. Li, C., Yang, Y., He, K., Lin, S., Hopcroft, J.E.: Single image reflection removal through cascaded refinement. In: Proc. of Computer Vision and Pattern Recognition (2020)
13. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proc. of Computer Vision and Pattern Recognition (2022)
14. Wan, R., Shi, B., Duan, L.Y., Tan, A.H., Gao, W., Kot, A.C.: Region-aware reflection removal with unified content and gradient priors. IEEE Transactions on Image Processing (2018)
15. Wan, R., Shi, B., Duan, L.Y., Tan, A.H., Kot, A.C.: CRRN: Multi-scale guided concurrent reflection removal network. In: Proc. of Computer Vision and Pattern Recognition (2018)
16. Wan, R., Shi, B., Li, H., Duan, L.Y., Kot, A.C.: Face image reflection removal. International Journal of Computer Vision (2021)
17. Wan, R., Shi, B., Li, H., Duan, L.Y., Tan, A.H., Kot, A.C.: CoRRN: Cooperative reflection removal network. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
18. Wan, R., Shi, B., Li, H., Hong, Y., Duan, L.Y., Kot, A.C.: Benchmarking single-image reflection removal algorithms. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
19. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution (2024)
20. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers (2003)
21. Wei, K., Yang, J., Fu, Y., Wipf, D., Huang, H.: Single image reflection removal exploiting misaligned training data and network enhancements. In: Proc. of Computer Vision and Pattern Recognition (2019)
22. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proc. of International Conference on Computer Vision (2023)
23. Zhang, X., Ng, R., Chen, Q.: Single image reflection separation with perceptual losses. In: Proc. of Computer Vision and Pattern Recognition (2018)
24. Zhong, H., Hong, Y., Weng, S., Liang, J., Shi, B.: Language-guided image reflection separation. arXiv preprint arXiv:2402.11874 (2024)